Răzvan V. Chereji · Alexandre V. Morozov

Statistical Mechanics of Nucleosomes Constrained by Higher-Order Chromatin Structure

May 24, 2011

Abstract Eukaryotic DNA is packaged into chromatin: one-dimensional arrays of nucleosomes separated by stretches of linker DNA are folded into 30-nm chromatin fibers which in turn form higher-order structures [1]. Each nucleosome, the fundamental unit of chromatin, has 147 base pairs (bp) of DNA wrapped around a histone octamer [2]. In order to describe how chromatin fiber formation affects nucleosome positioning and energetics, we have developed a thermodynamic model of finite-size particles with effective nearest-neighbor interactions and arbitrary DNA-binding energies. We show that both one- and two-body interactions can be accurately extracted from one-particle density profiles based on high-throughput maps of *in vitro* or *in vivo* nucleosome positions. Although a simpler approach that neglects two-body interactions (even if they are in fact present in the system) can be used to predict sequence determinants of nucleosome positions, the full theory is required to disentangle one- and two-body effects. Finally, we construct a minimal model in which nucleosomes are positioned by steric exclusion and two-body interactions rather than intrinsic histone-DNA sequence preferences. The model accurately reproduces nucleosome occupancy patterns observed over transcribed regions in living cells.

Keywords Chromatin structure · Nucleosome positioning · One-dimensional classical fluid of interacting particles

**PACS** 87.18.Wd · 87.80.St · 05.20.Jj

1 Introduction

Eukaryotic DNA in packaged into nucleosomes. Each nucleosome consists of a 147 bp-long DNA segment wrapped around a histone octamer in ~ 1.7 turns of a left-handed superhelix [2]. In addition to its primary function of DNA compaction, chromatin modulates DNA accessibility to transcription factors and other molecular machines, affecting gene transcription as well as DNA repair, maintenance, and replication. In *S. cerevisiae*, arrays of nucleosomes cover $\sim 80\%$ of genomic DNA, exerting profound influence on gene regulatory programs [3,4,5,6].

Because the free energy of bending a DNA segment into a superhelix will vary depending on its nucleotide sequence and composition [7,8], nucleosomes exhibit a range of *in vitro* formation energies [9,10] (although almost any DNA sequence can be packaged into a nucleosome). Recent work

R.V. Chereji
Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854-8019, USA E-mail: rchereji@physics.rutgers.edu

A.V. Morozov
Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854-8019, USA
BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, NJ 08854-8019, USA
E-mail: morozov@physics.rutgers.edu
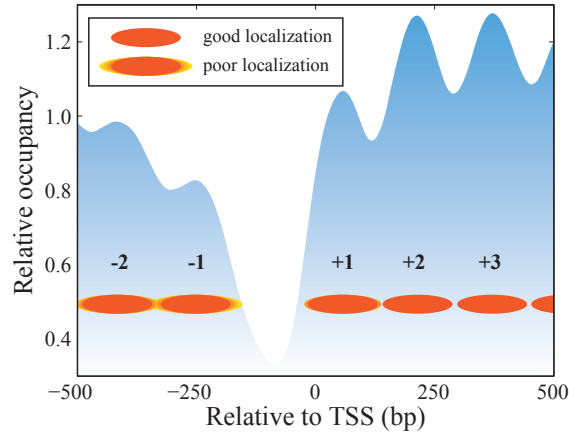
Fig. 1 (Color online) A typical nucleosome configuration near the transcription start site (TSS). The blue background represents *in vivo* nucleosome occupancy from Zawadzki et al. [21], averaged over all genes. The occupancy profile is normalized by its average in the $[-500, 500]$ bp window. The intensity of the blue color is proportional to the degree of nucleosome localization.

has clarified the role of sequence rules that influence nucleosome positioning: genome-wide *in vitro* reconstitution experiments have confirmed that nucleosome architecture over promoters and genes is partially established by DNA sequence, mostly as a result of nucleosome depletion from A/T-rich, nucleosome-disfavoring sequences on both ends of the transcript [11,12,13]. However, even if genomic DNA from *S. cerevisiae* is mixed with histones in a 1:1 mass ratio (leading to the maximum nucleosome occupancy of 0.82 which is close to the *in vivo* value [13]), nucleosomes are not strongly localized and, on average, nucleosome occupancy is just $\sim 20-30\%$ lower over nucleosome-depleted regions (NDRs) compared to the mean occupancy in a window which includes both the coding region and adjacent sequence. The absence of nucleosome localization *in vitro* and shallow NDRs indicate that the absolute magnitude of intrinsic histone-DNA interactions is less than 1 k_BT.

In vivo, 5' and 3' NDRs flanking the transcript are much more pronounced ($\sim 60-70\%$ occupancy depletion on average with respect to the mean [12,14,15,16]), establishing a striking pattern of nucleosome localization over genic regions simply due to steric exclusion which causes nucleosomes to "phase off" potential barriers [17] (Fig. 1). Although the exact nature of these *in vivo* barriers is unknown and may vary between cell types and environmental conditions, they are likely established through a combined action of RNA polymerase, ATP-dependent chromatin remodeling enzymes and DNA-binding proteins [18,19,20].

Nucleosome positions and formation energies can be predicted using a thermodynamic model which takes intrinsic histone-DNA sequence preferences and one-dimensional steric exclusion into account [22]. In this approach, sequence determinants of nucleosome energetics are inferred directly from experimentally available nucleosome occupancy profiles. The profiles are obtained by isolating and sequencing mononucleosomal DNA on a large scale, followed by mapping nucleosomal sequence reads to the reference genome [23]. However, structural regularity of the chromatin fiber imposes additional constraints on nucleosome positions [24,25]: linkers between neighboring nucleosomes become preferentially discretized with the $10-11$ bp periodicity of DNA helical twist [26]. The discretization is required to avoid steric clashes caused by the nucleosome rotating with respect to the linker DNA axis as the linker increases in length [24], and more generally to maintain a regular pattern of protein-protein and protein-DNA contacts in the chromatin fiber [25]. Indeed, adding a short DNA segment to the linker will rotate the nucleosome with respect to the rest of the fiber, causing disruption of its periodic structure. The disruption is minimized if the length of the extra segment is a multiple of $10-11$ bp, which brings the nucleosome into an equivalent rotational position.

We have recently developed a rigorous approach in which linker length discretization is described by nearest-neighbor two-body interactions in a system of non-overlapping finite-size particles [27] (Fig. 2). We have shown that it is possible to infer one-body energies given by intrinsic histone-DNA interactions simultaneously with two-body energies caused by chromatin fiber formation. The two-body potential
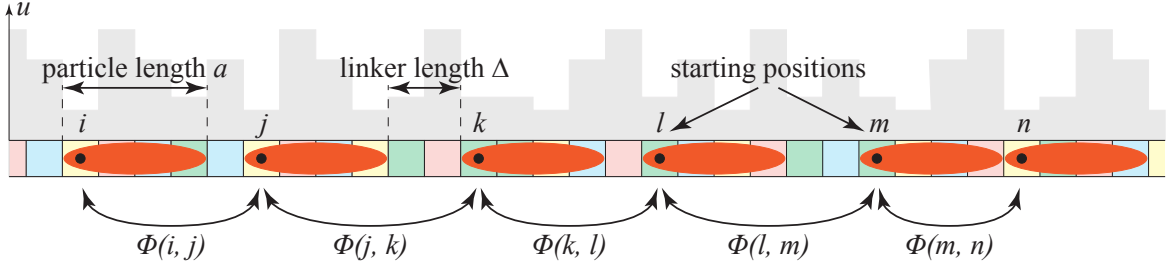
Fig. 2 (Color online) A typical configuration of 6 nucleosomes. Each nucleosome covers DNA sequence (represented by colored boxes) which gives the one-body energy of the nucleosome, u. The one-body energy is represented by gray bars. For simplicity, in this toy model, the one-body energy is entirely determined by the base pair located at the starting position of the nucleosome. In more realistic scenarios one-body energy is a function of the entire sequence occupied by the nucleosome. The two-body interaction $\Phi(i, j)$ acts only between neighboring nucleosomes, with two indices i and j representing their starting positions.

can be deduced even in the presence of one-body energies related to rotational positioning of the nucleosome [16,23,13], which have the same $10 - 11$ bp helical twist periodicity. We have predicted the two-body interaction from high-throughput maps of nucleosome positions on the *S. cerevisiae* genome, and demonstrated its essential role in forming nucleosome occupancy patterns over genic regions.

Here we present a detailed account of our theoretical framework. We also show that sequence determinants of nucleosome positioning can be obtained with an interaction-free model [22], even if in reality the two-body potential and histone-DNA interactions have comparable magnitudes. To this end, we develop a minimally constrained sequence-specific model of nucleosome energetics in which the same energies are assigned to mono- and dinucleotides regardless of their exact position within the 147 bp nucleosomal site [22]. However, only by taking into account the two-body interaction can we make a clear distinction between the two types of energies which dictate nucleosome positioning. Finally, we build a minimal model in which *in vivo* nucleosomes are positioned solely by potential barriers located at each end of the transcript. Without invoking explicit sequence specificity, the model successfully reproduces nucleosome occupancy patterns observed *in vivo* in *S. cerevisiae*. In contrast, sequence-dependent models can only capture liquid-like, delocalized behavior observed with *in vitro* nucleosomes [12,13]. By combining the minimal model with sequence-specific nucleosome energies, we estimate that intrinsic histone-DNA interactions contribute $< 30\%$ to the height of the *in vivo* potential barriers.

2 Theory

2.1 Energetics of one-dimensional hard rods with nearest-neighbor interactions

We consider the problem of interacting hard rods of length $a = 147$ bp confined to a one-dimensional lattice of length L bp (the length of the DNA segment) (Fig. 2). Let $u(k)$ be the external potential energy of a particle that occupies positions k through $k + a - 1$ on the DNA (the one-body energy), and let $\Phi(k, l)$ be the two-body interaction between a pair of nearest-neighbor particles with starting positions k and l, respectively. Here $u(k)$ describes intrinsic histone-DNA interactions, while $\Phi(k, l)$ accounts for the effects of chromatin structure. We assume that the DNA segment is surrounded by impenetrable walls, so that

$$u(0) = u(L - a + 2) = u(L - a + 3) = \ldots = u(L) = \infty.$$

Moreover, particle overlaps are not allowed and the two-body potential is short-range as in the Takahashi hard-rod model [28],

$$\Phi(k, l) = \begin{cases} \infty & \text{if } l < k + a, \\ 0 & \text{if } l \geq k + 2a. \end{cases}$$

The canonical partition function for a fixed number of particles N is given by

$$Q_N = \sum_{i_m + a \leq i_{m+1}} e^{-\beta u(i_1)} e^{-\beta \Phi(i_1, i_2)} e^{-\beta u(i_2)} \ldots e^{-\beta u(i_{N-1})} e^{-\beta \Phi(i_{N-1}, i_N)} e^{-\beta u(i_N)}, \tag{1}$$

where $\beta = 1/k_B T$ is the inverse temperature.

Let us introduce two $l_{\max} \times l_{\max}$ matrices, where $l_{\max} = L - a + 1$ is the rightmost starting position of a particle of length a:

$$\langle k|w|l\rangle = \begin{cases} e^{-\beta\Phi(k,l)} & \text{if } l \geq k + a, \\ 0 & \text{otherwise,} \end{cases}$$

$$\langle k|e|l\rangle = e^{-\beta u(k)}\delta_{k,l}.$$

Here $\delta_{k,l}$ is the Kronecker delta symbol, and $\langle k|M|l\rangle$ represents the element of matrix M in row k and column l (in Dirac notation). $|l\rangle$ is a column vector of dimension $l_{\max}$ with 1 at position l and 0 everywhere else, and $\langle k|$ is a row vector with 1 at position k.

Defining $|J\rangle = \sum_{l=1}^{l_{\max}} |l\rangle$ (a vector with 1 at every position), we rewrite Eq. (1) as

$$Q_N = \begin{cases} \langle J|(ew)^{N-1}e|J\rangle & \text{if } N \geq 1, \\ 1 & \text{if } N = 0. \end{cases}$$

The grand-canonical partition function is then given by

$$Z = \sum_{N=0}^{N_{\max}} e^{\beta N\mu}Q_N = 1 + \langle J|(I - zw)^{-1}z|J\rangle, \tag{2}$$

where μ is the chemical potential, $N_{\max} = \lfloor \frac{L}{a} \rfloor$ is the maximum number of particles that can fit on L bp, I is the identity matrix, and $\langle k|z|l\rangle = e^{\beta[\mu - u(k)]}\delta_{k,l}$.

The s-particle distribution functions are defined as

$$n_s(i_1, \ldots, i_s) \equiv \frac{\xi(i_1) \ldots \xi(i_s)}{Z} \frac{\delta^s Z}{\delta\xi(i_1) \ldots \delta\xi(i_s)},$$

where $\xi(i) = e^{\beta[\mu - u(i)]}$ (see the chapter by Stell in [29]). The one-particle distribution function is

$$n(i) = \frac{1}{Z}\langle J|(I - zw)^{-1}|i\rangle\langle i|z|i\rangle\langle i|(I - wz)^{-1}|J\rangle, \tag{3}$$

and the two-particle, nearest-neighbor distribution function is

$$\overline{n}_2(i,j) = \frac{1}{Z}\langle J|(I - zw)^{-1}|i\rangle\langle i|zwz|j\rangle\langle j|(I - wz)^{-1}|J\rangle. \tag{4}$$

These relations are easy to understand. To find the probability of starting a particle at position i [Eq. (3)], we have to add the statistical weights of all configurations that contain a particle at that position, and divide the resulting sum by the partition function. Similarly, to find the probability of having a pair of nearest-neighbor particles with the starting positions i and j [Eq. (4)], we need to sum the statistical weights of all configurations that contain that pair of particles.

Note that for short distances $j - i < 2a$, $\overline{n}_2(i,j)$ is identical to the unrestricted two-particle distribution $n_2(i,j)$, because there is not enough space to put another particle between the two particles at i and j. In this paper we restrict ourselves to $j - i < 2a$, since we are interested in short-range interactions between nearest-neighbor pairs of particles.

In many cases of interest the energetics of the system is unknown but the s-particle distributions are available from experiment. Therefore, we wish to find the unknown energies u and Φ from n and $\overline{n}_2$ by inverting Eqs. (3) and (4). Let us define two matrices: $\langle i|N|j\rangle = n(i)\delta_{i,j}$ and $\langle i|N_2|j\rangle = \overline{n}_2(i,j)$. Using these matrices and Eq. (2), we have

$$\langle J|(I - N_2 N^{-1})N|J\rangle = \frac{1}{Z}\langle J|(I - zw)^{-1}z|J\rangle = \frac{Z - 1}{Z},$$

so that

$$Z = \frac{1}{1 - \langle J|(I - N_2 N^{-1})N|J\rangle}. \tag{5}$$

After some matrix multiplications we obtain the following two equalities:

$$\langle J|I - N_2 N^{-1}|k\rangle = \frac{1}{\langle J|(I - zw)^{-1}|k\rangle},$$

(6)

and

$$\langle k|I - N^{-1}N_2|J\rangle = \frac{1}{\langle k|(I - wz)^{-1}|J\rangle}.$$

(7)

Substituting Eqs. (5), (6) and (7) into Eqs. (3) and (4), we derive the exact expressions for one-body energies and two-body interactions [30,31]:

$$-\beta\left[u(k) - \mu\right] = \ln\left(\frac{\langle J|I - N_2 N^{-1}|k\rangle\langle k|N|k\rangle\langle k|I - N^{-1}N_2|J\rangle}{1 - \langle J|(I - N_2 N^{-1})N|J\rangle}\right),$$

(8)

$$-\beta\Phi(k,l) = \ln\left(\frac{\langle k|N^{-1}N_2 N^{-1}|l\rangle\left[1 - \langle J|(I - N_2 N^{-1})N|J\rangle\right]}{\langle k|I - N^{-1}N_2|J\rangle\langle J|I - N_2 N^{-1}|l\rangle}\right).$$

(9)

If the two-body interactions other than the steric exclusion are neglected, then $\langle k|w|l\rangle \to \Theta(l-k-a)$, where $\Theta(l-k-a)$ is the Heaviside step function (1 if $l \geq k+a$, and 0 otherwise). In this case we have:

$$\langle J|(I - zw)^{-1}|i\rangle = \langle J|I + zw + (zw)^2 + \dots|i\rangle = Z_{i-1}^f,$$
$$\langle i|(I - wz)^{-1}|J\rangle = \langle i|I + wz + (wz)^2 + \dots|J\rangle = Z_{i+a}^r,$$

where Z_i^f and Z_i^r are partial statistical sums which can be efficiently computed by iteration in a forward or reverse direction [8,22]. Note that $Z = Z_1^r = Z_{L-a+1}^f$. The partial statistical sums account for the contributions from all possible configurations of particles confined to the boxes $[1,i]$ and $[i,L]$, respectively. It can be shown that [22]:

$$Z_i^f = \prod_{j=1}^{i} \frac{1 - O(j+1) + n(j+1)}{1 - O(j)},$$

$$Z_i^r = \prod_{j=i}^{L} \frac{1 - O(j) + n(j)}{1 - O(j)},$$

where $O(i)$ is the particle occupancy of bp i $\left[O(i) = \sum_{j=i-a+1}^{i} n(j)\right]$.

Using Eq. (3), we reproduce the previous result from [22] which can be employed to find one-body energies from one-particle distribution in the case of hard-core interactions alone,

$$-\beta\left[u^0(i) - \mu\right] = \ln\left[\frac{n(i)}{1 - O(i) + n(i)}\right] + \ln\left[\prod_{j=i}^{i+a-1}\frac{1 - O(j) + n(j)}{1 - O(j)}\right].$$

(10)

2.2 Predicting two-body interactions from one-particle distribution

As shown above, there is a one-to-one correspondence between one-body energies u and two-body interactions Φ on one hand, and particle distributions n and $\overline{n}_2$ on the other. Thus, if n and $\overline{n}_2$ are known, u and Φ can be inferred exactly, and vice versa. However, in many situations the two-particle distribution is not directly available from experiments. For example, high-throughput nucleosome maps simultaneously report nucleosome positions from many cells, effectively yielding a probabilistic description of the one-particle distribution n. Because of this averaging over single-cell configurations, information about the pair density profile $\overline{n}_2$ cannot be extracted directly. Nonetheless, if the two-body interactions are sufficiently strong, the one-particle distribution profile n can be used to obtain information about Φ.

Let us introduce the dimensionless pair distribution

$$g(i,j) = \frac{n_2(i,j)}{n(i)n(j)}. \tag{11}$$

Note that $g(i,j) = \bar{n}_2(i,j)/[n(i)n(j)]$ for short distances $j - i < 2a$, and that $g(i,j) = g(j-i)$ in a homogeneous system. We start with a homogeneous system of N hard-rods which interact through an arbitrary nearest-neighbor potential Φ, and then develop an approximation for the inhomogeneous case. In a translation-invariant continuous system with nearest-neighbor interactions of arbitrary strength, $e^{-\beta\Phi(d)} = Ce^{\alpha d}g(d)$, where C and α are constants [28,32,33,34]. The result can also be proved for a lattice fluid of hard rods, as shown below.

Proof Consider a system of N particles distributed on a segment of length L bp. We assume that the particles interact with each other through short-range nearest-neighbor interactions (which include steric exclusion if the particles have a finite size of a bp), and the total interaction energy is

$$U(x_1, x_2, \ldots, x_N) = \Phi(x_2 - x_1) + \Phi(x_3 - x_2) + \ldots + \Phi(x_N - x_{N-1}) + U_b,$$

where U_b is the boundary term which describes interaction between the walls and the first and last particles.

For simplicity let us assume that the boundary conditions are enforced by two additional particles of the same kind fixed at $x = 0$ and $x = L$, so that $U_b = \Phi(x_1) + \Phi(L - x_N)$. The exact form of boundary conditions is not essential in the thermodynamic limit. The canonical partition function of this system of N particles is

$$Q_N(L) = \sum_{0 \leq x_1 \leq x_2 \leq \ldots \leq x_N \leq L} e^{-\beta\Phi(x_1-0)} e^{-\beta\Phi(x_2-x_1)} \ldots e^{-\beta\Phi(L-x_N)}$$

$$= \sum_{x_N=0}^{L} \sum_{x_{N-1}=0}^{x_N} \ldots \sum_{x_1=0}^{x_2} e^{-\beta\Phi(x_1-0)} e^{-\beta\Phi(x_2-x_1)} \ldots e^{-\beta\Phi(L-x_N)}$$

Denoting $f(x) \equiv e^{-\beta\Phi(x)}$, we obtain

$$Q_N(L) = \sum_{x_N=0}^{L} \sum_{x_{N-1}=0}^{x_N} \ldots \sum_{x_1=0}^{x_2} f(x_1 - 0)f(x_2 - x_1)\ldots f(L - x_N).$$

Note that this represents the convolution of $N + 1$ functions f,

$$Q_N(L) = (\underbrace{f * f * \ldots * f}_{N+1 \text{ functions}})(L).$$

The partition function can be computed using the z transform method. Let $\widetilde{Q}(z)$ be the z transform of $Q_N(L)$,

$$\widetilde{Q}(z) = \sum_{n=0}^{\infty} Q_N(n)z^{-n}.$$

From the convolution theorem we have that

$$\widetilde{Q}(z) = \left[\widetilde{F}(z)\right]^{N+1},$$

where $\widetilde{F}(z)$ is the z transform of $f(n)$,

$$\widetilde{F}(z) = \sum_{n=0}^{\infty} f(n)z^{-n}.$$

The partition function can be recovered using the inverse z transform,

$$Q_N(L) = \frac{1}{2\pi i} \oint_{\Gamma} \left[\widetilde{F}(z)\right]^{N+1} z^{L-1} z.$$

The contour of integration Γ is any simple closed curve enclosing $|z| = R$, $|z| > R$ being the region of convergence.

Let us define $h(z) = (N+1)\ln \widetilde{F}(z) + (L-1)\ln z$. With this notation

$$Q_N(L) = \frac{1}{2\pi i} \oint_\Gamma e^{h(z)}\frac{dz}{z}.$$

This integral can be computed by the saddle point method [34]. Expanding $h(z)$ around the saddle point z_0, we obtain

$$Q_N(L) \approx e^{h(z_0)}\frac{1}{2\pi i}\int e^{\frac{1}{2}h''(z_0)(z-z_0)^2}\frac{dz}{z}.$$

Integration along the path of steepest descent yields a contribution from the Gaussian integral of order $O\left([h''(z_0)]^{-1/2}\right) = O\left(N^{-1/2}\right)$. Since we need $\ln Q_N(L)$ in order to compute the macroscopic quantities, and in the thermodynamic limit the terms of order $O(\ln N)$ are not important, we can approximate the partition function as

$$Q_N(L) \approx e^{h(z_0)} \approx z_0^L \left[\widetilde{F}(z_0)\right]^N, \tag{12}$$

where z_0 is the saddle point, given by

$$\left.\frac{dh}{dz}\right|_{z=z_0} \approx \frac{L}{z_0} + N\frac{\widetilde{F}'(z_0)}{\widetilde{F}(z_0)} = 0. \tag{13}$$

We can compute the chemical potential for the interacting hard rods by taking the derivative of the free energy F with respect to the number of particles in the system

$$\mu = \frac{\partial F}{\partial N} = -k_B T\frac{\partial \ln Q_N}{\partial N} = -k_B T\ln \widetilde{F}(z_0). \tag{14}$$

The pressure of the gas is given by the derivative of the free energy with respect to the length of the system. Let's denote the length of a base pair by b, such that the real length of the system is Lb. We obtain

$$p = -\frac{1}{b}\frac{\partial F}{\partial L} = \frac{k_B T}{b}\frac{\partial \ln Q_N(L)}{\partial L} = \frac{k_B T}{b}\ln z_0, \tag{15}$$

and from Eq. (12) we obtain

$$Q_N(L) = \left(e^{\beta pb}\right)^L\left[\widetilde{F}\left(e^{\beta pb}\right)\right]^N.$$

We use this result to compute the conditional probability of finding an adjacent particle at a distance d from the center of a fixed particle [32,35]:

$$P(d) \equiv \mathrm{Prob}(x_N = L - d|x_{N+1} = L)$$
$$= \frac{1}{Q_N}\sum_{x_{N-1}=0}^{x_N}\ldots\sum_{x_1=0}^{x_2}f(x_1)\ldots f((L-d)-x_{N-1})f(L-(L-d))$$
$$= f(d)\frac{Q_{N-1}(L-d)}{Q_N(L)}$$
$$= e^{-\beta\Phi(d)}\frac{\left(e^{\beta pb}\right)^{-d}}{\widetilde{F}(e^{\beta pb})}.$$
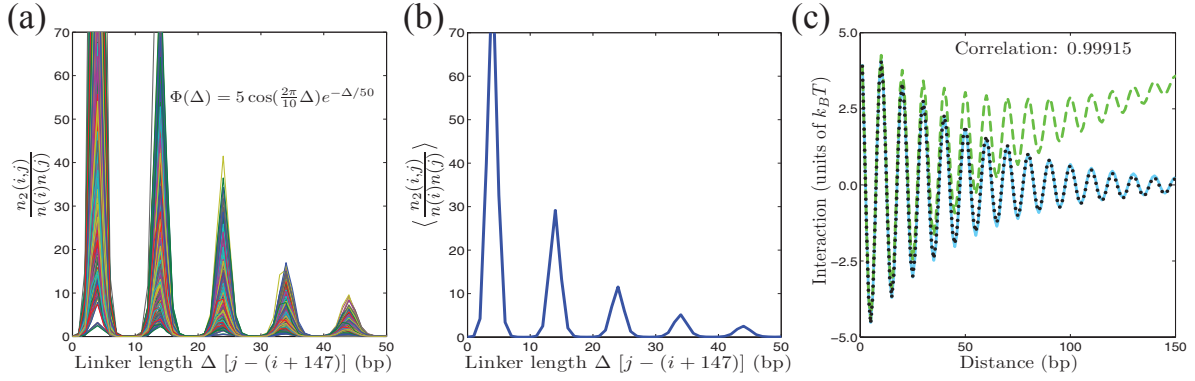
For $d < 2a$ the pair distribution becomes

$$g(d) = \frac{\overline{n}_2(i,i+d)}{n(i)n(i+d)} = \frac{\rho P(d)}{\rho^2} \propto e^{-\beta pbd}e^{-\beta\Phi(d)}.$$

Thus in the homogeneous system the interaction between the particles has the form

$$e^{-\beta\Phi(d)} = Ce^{\alpha d}g(d), \tag{16}$$

where $\alpha = \beta pb$ and C is a normalization constant. $\quad\square$

Fig. 3 (Color online) (a) $g(i,j) = n_2(i,j)/n(i)n(j)$ is plotted for a representative subset of all initial positions i in a 10^4 bp DNA segment. The one-body energies are randomly sampled from a Gaussian distribution with a mean of 2.5 k_BT and a standard deviation of 0.2 k_BT, and 9 potential wells of depth 5 k_BT are added at $1, 2, \ldots, 9 \times 10^3$ bp to model a strongly inhomogeneous system. $n_2(i,j)$ and $n(i)$ are computed from one- and two-body energies using Eqs. (3) and (4). (b) $P_{\text{linker}}(\Delta)$, obtained by averaging $g(i,j)$ over all initial positions i. Note that $\Delta = j - (i + 147)$ represents the linker length between the two nucleosomes with starting positions i and j, respectively. (c) Exact (solid blue line) and predicted (dotted black line) two-body interactions. The predicted interaction was computed from the $-\ln(P_{\text{linker}})$ curve (dashed green line) using Eq. (17).

In a more general case, the external potential breaks translational invariance, making g dependent on the absolute position of the first particle. However, if the two-body interaction Φ is translationally invariant, a good approximation is provided by replacing g with $P_{\text{linker}}(\Delta) = \langle g(i, i+a+\Delta) \rangle_i$ (averaged over all initial positions i) in Eq. (16) [27],

$$-\beta\Phi(i,j) \approx \ln\left[P_{\text{linker}}(j - (i + a))\right] + \alpha(j - i) + \ln C \qquad (17)$$

The constants C and α are uniquely determined by the asymptotic condition $\lim_{(j-i)\to\infty} \Phi(i,j) = 0$. Eq. (17) provides an ansatz for reconstructing Φ from $P_{\text{linker}}(\Delta) = \langle \overline{n}_2(i, i + a + \Delta)/[n(i)n(i + \Delta)] \rangle_i$.

Fig. 3 shows a numerical test of this ansatz on a 10 kbp DNA segment. We construct a random one-body energy landscape and simulate strong inhomogeneity by positioning 9 potential wells with depth of $5k_BT$ at $1, 2, \ldots, 9$ kbp on the landscape. The model interaction between a pair of particles separated by a linker of length Δ is $\Phi(\Delta) = 5 \cos\left(\frac{2\pi}{10}\Delta\right) e^{-\Delta/50}$ (in units of k_BT). We use the one-body energies and the two-body potential as inputs to Eqs. (3) and (4), computing the dimensionless pair distribution function. The pair distribution varies significantly from bp to bp [Fig. 3(a)], as can be expected in a system with one- and two-body energies of comparable magnitude. Following our prescription, we compute P_{linker} by averaging over all the curves in Fig. 3(a) [Fig. 3(b)], and employ Eq. (17) to infer Φ [Fig. 3(c)]. The correlation coefficient between predicted and exact two-body interactions is greater than 0.999.

If a direct measurement of the pair distribution $\overline{n}_2$ is not available, P_{linker} needs to be estimated empirically from the n profile. Each nucleosome positioning data set consists of the histogram of the number of nucleosomes starting at each genomic bp i. We preprocess these data by removing all counts of height 1 from the histogram and smoothing the remaining counts with a $\sigma = 2$ Gaussian kernel. Next, we compute $n(i)$ by rescaling the smoothed profile so that the maximum occupancy for each chromosome is 1. Finally, we identify all local maxima on the n profile and assume that they mark prevalent nucleosome positions. Specifically, for each maximum at bp i we find subsequent maxima at positions $i + 146 < j_1 < j_2 < j_3 < \ldots$ in the 50 bp window. To each pair of maxima (i, j_1), $(i, j_2), \ldots$ we assign the probability that they represent neighboring nucleosomes: $n(i)n(j_1)$, $n(i)[1-n(j_1)]n(j_2)$, and so on. We sum the probabilities over all initial positions i and normalize, producing an empirical estimate of P_{linker}.

2.3 Comparison between lattice and continuous one-dimensional fluids

The formalism presented in the previous subsection can be used to obtain the equation of state and chemical potential for any generic interaction $\Phi(x)$. Let us consider two simple cases: the ideal gas and the Tonks lattice gas, which is characterized by the hard-core interaction

$$\Phi(x) = \begin{cases} \infty & \text{if } x < a, \\ 0 & \text{if } x \geq a. \end{cases}$$

For the ideal gas, the z transform of $e^{-\beta\Phi}$ and the saddle point z_0 [obtained from Eq. (13)] are given by:

$$\widetilde{F}(z) = \frac{z}{z-1},$$
$$z_0 = \frac{L+N}{N}.$$

Using these expressions, we can compute the logarithm of the partition function [Eq. (12)]

$$\ln Q_N(L) = L \ln\left(\frac{L+N}{L}\right) + N \ln\left(\frac{L+N}{N}\right) + O(\ln N),$$

which gives the pressure and the chemical potential for the ideal lattice gas:

$$\beta p_1^{\text{id}} = \frac{1}{b} \ln\left(1 + \frac{N}{L}\right), \tag{18}$$

$$\beta \mu_1^{\text{id}} = \ln\left(\frac{N}{L+N}\right). \tag{19}$$

In the case of the Tonks lattice gas,

$$\widetilde{F}(z) = \frac{z^{1-a}}{z-1},$$
$$z_0 = \frac{L - Na + N}{L - Na}.$$

The pressure and the chemical potential are then given by:

$$\beta p_1^{\text{T}} = \frac{1}{b} \ln\left(1 + \frac{N}{L-Na}\right), \tag{20}$$

$$\beta \mu_1^{\text{T}} = a \ln\left(\frac{L-Na+N}{L-Na}\right) + \ln\left(\frac{N}{L-Na+N}\right). \tag{21}$$

It is useful to compare these results with the corresponding results for continuous one-dimensional gases. Denoting the physical length of the particles by $\mathcal{A}$, the length of the box by $\mathcal{L}$, and the Laplace transform of $e^{-\beta\Phi}$ by

$$\varphi(s) = \int_0^\infty s^{-sx} e^{-\beta\Phi(x)} x,$$

we obtain the canonical partition function as the inverse Laplace transform

$$Q_{\text{cont}}(N, \mathcal{L}, T) = \frac{1}{\lambda(T)^N} \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{s\mathcal{L}} [\varphi(s)]^N s \approx e^{s_0\mathcal{L}} \left[\frac{\varphi(s_0)}{\lambda(T)}\right]^N,$$

where $\lambda(T) = h/\sqrt{2\pi m k_B T}$ is the thermal de Broglie wavelength, and s_0 is the saddle point given by the equation

$$\mathcal{L} + N \frac{\varphi'(s_0)}{\varphi(s_0)} = 0.$$

Using the previous two equations, we find for the ideal gas $\left[\varphi(s) = \frac{1}{s}, \text{ and } s_0 = \frac{N}{\mathcal{L}}\right]$:

$$\beta p_{\mathrm{c}}^{\mathrm{id}} = \frac{\partial \ln Q_{\mathrm{cont}}}{\partial \mathcal{L}} = \frac{N}{\mathcal{L}}, \tag{22}$$

$$\beta \mu_{\mathrm{c}}^{\mathrm{id}} = -\frac{\partial \ln Q_{\mathrm{cont}}}{\partial N} = \ln \frac{N\lambda(T)}{\mathcal{L}}. \tag{23}$$

Similarly, for the Tonks gas we obtain $\left[\varphi(s) = \frac{e^{-\mathcal{A}s}}{s}, \text{ and } s_0 = \frac{N}{\mathcal{L} - N\mathcal{A}}\right]$:

$$\beta p_{\mathrm{c}}^{\mathrm{T}} = \frac{N}{\mathcal{L} - N\mathcal{A}}, \tag{24}$$

$$\beta \mu_{\mathrm{c}}^{\mathrm{T}} = \frac{N\mathcal{A}}{\mathcal{L} - N\mathcal{A}} + \ln \left[\frac{N\lambda(T)}{\mathcal{L} - N\mathcal{A}}\right]. \tag{25}$$

To compare continuous and discrete results, we let the lattice constant b approach 0, while keeping the particle size ($\mathcal{A} = ab$) and the box size ($\mathcal{L} = Lb$) finite. We obtain:

$$\lim_{b \to 0} \beta p_{\mathrm{l}}^{\mathrm{id}} = \lim_{b \to 0} \frac{1}{b} \ln \left(1 + \frac{Nb}{\mathcal{L}}\right) = \beta p_{\mathrm{c}}^{\mathrm{id}},$$

$$\lim_{b \to 0} \beta p_{\mathrm{l}}^{\mathrm{T}} = \lim_{b \to 0} \frac{1}{b} \ln \left(1 + \frac{Nb}{\mathcal{L} - N\mathcal{A}}\right) = \beta p_{\mathrm{c}}^{\mathrm{T}}.$$

Similarly, the chemical potentials for the ideal and Tonks lattice gases become asymptotically, as $b \to 0$:

$$\beta \mu_{\mathrm{l}}^{\mathrm{id}} \sim \ln \frac{Nb}{\mathcal{L}},$$

$$\beta \mu_{\mathrm{l}}^{\mathrm{T}} \sim \frac{N\mathcal{A}}{\mathcal{L} - N\mathcal{A}} + \ln \left(\frac{Nb}{\mathcal{L} - N\mathcal{A}}\right).$$

These expressions are identical to the chemical potentials of the corresponding continuous gases [Eqs. (23) and (25)], with the length scale $\lambda(T)$ replaced by the typical length scale of the lattice, b.

2.4 Sequence-specific energy of nucleosome formation

We can extract a sequence-specific component of the one-body energy by using Eqs. (8) or (10) to compute $u - \mu$, estimating the chemical potential μ, and fitting the one-body energy u to a linear model which assigns energies to nucleotide words found within the $a = 147$ bp nucleosomal site. Assuming that the system is nearly homogeneous, we use Eqs. (14) or (21) to obtain the chemical potential of the lattice gas. After eliminating μ, we fit a linear model to one-body energies u. It was established in Ref. [22] that position-independent models in which the energy of the nucleotide word does not depend on its exact location within the nucleosome can be used to describe genome-wide nucleosome occupancies. Furthermore, an $N = 2$ position-independent model with just 13 fitting parameters performed as well as $N > 2$ models [here, N denotes the longest word (in bp) included into the model].

If both monomers and dimers contribute to the total one-body energy, the sequence-specific binding energy of a 147 bp-long nucleosomal site is given by

$$u^S = \sum_{\alpha} m_{\alpha} \epsilon_{\alpha} + \sum_{\alpha,\beta} m_{\alpha\beta} \epsilon_{\alpha\beta} + \epsilon_0, \tag{26}$$

where m_{α} is the number of nucleotides of type $\alpha \in \{A, C, G, T\}$, ϵ_{α} is the energy of the nucleotide α, and ϵ_0 is the overall sequence-independent offset. Similarly, $m_{\alpha\beta}$ is the number of dinucleotides of type $\alpha\beta$, and $\epsilon_{\alpha\beta}$ is the corresponding energy. In Ref. [22], word energies were constrained by $\sum_{\alpha} \epsilon_{\alpha} = \sum_{\alpha} \epsilon_{\alpha\beta} = \sum_{\beta} \epsilon_{\alpha\beta} = 0$, yielding a 13-parameter model. Here we develop an alternative

approach which does not impose any additional constraints beyond those caused by the fact that the number of mono- and dinucleotides in the 147 bp-long site is fixed.

We can express the nucleosome energies as $\mathbf{u} = \mathbf{M}\mathbf{x}$, or equivalently

$$\begin{pmatrix} u^S(1) \\ u^S(2) \\ \vdots \\ u^S(l_{max}) \end{pmatrix} = \begin{pmatrix} m_{1,1} & \cdots & m_{1,20} & 1 \\ m_{2,1} & \cdots & m_{2,20} & 1 \\ \vdots & & \vdots & \vdots \\ m_{l_{max},1} & \cdots & m_{l_{max},20} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_{21} \end{pmatrix},$$

where $u^S(i)$ is the sequence-specific energy of the nucleosome that covers the DNA sequence between bps i and $i+a-1$. $m_{i,1}, \ldots, m_{i,4}$ give the number of A, C, G and T nucleotides found in that sequence, and $m_{i,5}, \ldots, m_{i,20}$ give the number of dinucleotides $AA, AC, \ldots, TG, TT$. $l_{max} = L - a + 1$ is the maximum starting position for a nucleosome, and the set of parameters $x_1, \ldots x_{21}$ represents the 21 energies from Eq. (26): $x_1 = \epsilon_A$, $x_2 = \epsilon_C$, ..., $x_{20} = \epsilon_{TT}$, $x_{21} = \epsilon_0$.

Note that for any DNA sequence of length 147 bp,

$$m_{i,1} + m_{i,2} + m_{i,3} + m_{i,4} = 147 \, m_{i,21},$$
$$m_{i,5} + m_{i,6} + \ldots + m_{i,20} = 146 \, m_{i,21}.$$

This means that the rank of $\mathbf{M}$ is 19. For any linear operator ($\mathbf{M}$), the dimension of its domain (21 in our case) is equal to the sum of the dimensions of the image $\text{im}(\mathbf{M})$ in the energy space (19) and the kernel $\ker(\mathbf{M})$ in the parameter space (2).

It is easy to check that the vectors

$$\mathbf{x}^* = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ \vdots \\ 0 \\ -147 \end{pmatrix}, \quad \mathbf{x}^{**} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \\ 1 \\ -146 \end{pmatrix}$$

belong to $\ker(\mathbf{M})$, so that $\ker(\mathbf{M}) = \text{span}(\mathbf{x}^*, \mathbf{x}^{**})$.

Any vector of parameters $\mathbf{x}$ can be uniquely decomposed as $\mathbf{x} = \mathbf{x}_K + \mathbf{x}^\perp$, with $\mathbf{x}_K \in \ker(\mathbf{M})$ and $\mathbf{x}^\perp \in \ker(\mathbf{M})^\perp$, the subspace which is orthogonal to $\ker(\mathbf{M})$. Specifically, $\mathbf{x}^\perp = \mathbf{x} - (\mathbf{x}^*, \mathbf{x})\mathbf{x}^* - (\mathbf{x}^{**}, \mathbf{x})\mathbf{x}^{**}$. The component $\mathbf{x}_K$ does not contribute to the energy of the sequence because $\mathbf{M}\mathbf{x}_K$ is the null vector. The components of $\mathbf{x}^\perp$ satisfy the following relations:

$$(\mathbf{x}^\perp, \mathbf{x}^*) = 0 \Rightarrow \sum_{i=1}^{4} x_i^\perp - 147 \, x_{21}^\perp = 0,$$

$$(\mathbf{x}^\perp, \mathbf{x}^{**}) = 0 \Rightarrow \sum_{i=5}^{20} x_i^\perp - 146 \, x_{21}^\perp = 0.$$

Thus $\mathbf{x}^\perp$ has only 19 independent parameters, and

$$x_{21}^\perp = \frac{1}{147} \sum_{i=1}^{4} x_i^\perp,$$

$$x_{20}^\perp = 146 \, x_{21}^\perp - \sum_{i=5}^{19} x_i^\perp = \frac{146}{147} \sum_{i=1}^{4} x_i^\perp - \sum_{i=5}^{19} x_i^\perp$$

$$\Rightarrow \mathbf{x}^\perp = \begin{pmatrix} x_1^\perp \\ \vdots \\ x_{19}^\perp \\ \frac{146}{147}\sum_{i=1}^{4} x_i^\perp - \sum_{i=5}^{19} x_i^\perp \\ \frac{1}{147}\sum_{i=1}^{4} x_i^\perp \end{pmatrix} \tag{27}$$

In order to compare two different sets of 21 energies (e.g. fit on different genomes), we need to eliminate the components of these two vectors included in $\ker(\mathbf{M})$. The components from $\ker(\mathbf{M})^\perp$ will have 19 independent parameters and 2 redundant ones [Eq. (27)]. The projection of the energy vector on the $\mathrm{im}(\mathbf{M})$ hyperplane is unique, and there is a one-to-one correspondence between $\mathrm{im}(\mathbf{M})$ and the parameter subspace which is orthogonal to the kernel, $\ker(\mathbf{M})^\perp$. In this way, every set of fitted energies uniquely determines a set of parameters ($\mathbf{x}^\perp$) and a sequence-specific energy ($\mathbf{u} = \mathbf{M}\mathbf{x}^\perp$).

For the $N = 1$ model, $\ker(\mathbf{M})$ is spanned by a single vector

$$\mathbf{x}^* = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -147 \end{pmatrix},$$

and $\mathbf{x}^\perp$ has 4 relevant parameters and a redundant one

$$\mathbf{x}^\perp = \begin{pmatrix} x_1^\perp \\ x_2^\perp \\ x_3^\perp \\ x_4^\perp \\ \frac{1}{147}\sum_{i=1}^{4} x_i^\perp \end{pmatrix}.$$

Similarly, the $N = 3$ model has 85 $(4+16+64+1)$ fitting parameters and there are six independent constraints on the columns of $\mathbf{M}$, so that the rank of the operator $\mathbf{M}$ is 79. The kernel of $\mathbf{M}$ is spanned by six vectors, and the parameter subspace orthogonal to the kernel (which gives the sequence energy) is 79-dimensional. For the $N = 4$ and the $N = 5$ models, the total number of parameters is 341 and 1365, and the number of independent parameters is 319 and 1279, respectively.

When the $N = 2$, 21-parameter model described above is trained on the energies predicted by applying Eq. (10) to a large-scale map of nucleosomes reconstituted *in vitro* on yeast genomic DNA [13], it captures the same sequence determinants as our previously used 13-parameter model which employs additional constraints [22] ($r = 0.9967$ between the two sequence-specific energy profiles). However, the two approaches are not equivalent, since the 21-parameter model utilizes the maximum possible number of independent fitting parameters.

3 Results

3.1 Reconstructing nucleosome energetics in a model system

In the absence of nearest-neighbor interactions induced by chromatin structure, nucleosome formation *in vitro* is fully controlled by DNA sequence and steric exclusion. In this case efficient procedures are available for reconstructing nucleosome positions from formation energies [8,36] and for inferring nucleosome energetics from experimentally available probability and occupancy profiles [Eq. (10)] [22]. However, this simple approach may lead to errors if the two-body interactions are in fact present in the system. Furthermore, many factors other than DNA sequence can affect nucleosome positioning *in vivo*, including chromatin remodeling enzymes, non-histone DNA-binding factors, and components of transcriptional machinery [18,19,20]. These influences are expected to create potential barriers which prevent nucleosomes from forming in certain regions, and potential wells which localize nucleosomes through favorable contacts between histones and other proteins. These effects will be lost if a purely sequence-specific model is fit to the nucleosome positioning data.

We use a simple model system to illustrate the errors caused by neglecting higher-order chromatin structure and *in vivo* potentials (Fig. 4). We have generated a random 10^4 bp DNA fragment and computed one-body sequence-dependent energies using the 21-parameter $N = 2$ position-independent model (section 2.4). The sequence-specific word energies for the model were randomly sampled from a $[-0.02 \ k_B T, 0.02 \ k_B T]$ uniform distribution. Fig. 4(a) shows sequence-dependent nucleosome energies in a representative 500 bp window (blue solid line). The window also includes one of the 3 $k_B T$ wells placed every 2000 bp throughout the sequence to model *in vivo* effects; the total one-body energy is shown as a green dash-dot-dot line.

The total energy can be used together with the two-body interaction shown in Fig. 4(b) (blue solid line) to construct the exact one-body density profile $n(i)$ and the corresponding nucleosome occupancy for the DNA segment [Fig. 4(c), green dash-dot-dot line]. If we now use Eq. (10) which neglects two-body interactions to reconstruct one-body energies from $n(i)$, the predicted energy profile captures the potential wells and the sequence-specific component but also displays spurious 10 bp oscillations caused by the "leakage" of the two-body potential Φ into one-body energetics [Fig. 4(a), red dashed line]. In addition, the whole landscape is shifted downward because favorable two-body interactions are missing from the model. The "leaked" oscillations and the *in vivo* wells have no relation to sequence and can be removed by fitting either the 13- or the 21-parameter model to the prediction [Fig. 4(a), light blue dash-dot line]. The two predicted energy profiles are highly correlated with each other ($r = 0.9993$) and with the exact profile ($r = 0.9913$ for the 13-parameter model, $r = 0.9915$ for the 21-parameter model), indicating that the sequence-specific component can be extracted even if the two-body interactions are not handled correctly.

Predicting occupancies from the energy profiles constructed under the $\Phi = 0$ assumption causes discrepancies with the exact result, shown in Fig. 4(c) as a green dash-dot-dot line. For example, using the one-body energies predicted with Eq. (10) [Fig. 4(a), red dashed line] and the two-body potential predicted with Eq. (17) [Fig. 4(b), green dashed line] gives an occupancy profile with higher average occupancy, sharp peaks, and enhanced 10 bp oscillations compared to the exact landscape [Fig. 4(c), red dashed line]. This is not unexpected because the two-body potential is both imprinted in the one-body profile and included explicitly. In contrast, if Φ is neglected at this stage as well, the exact occupancy can be restored from u^0 [Eq. (10) and its inverse do not entail any information loss], but the origin of various contributions remains unclear as they are all lumped into the one-body landscape.

When the 21-parameter model is fit to the $\Phi = 0$ profile [Fig. 4(a), light blue dash-dot line] and combined with the predicted Φ [Fig. 4(b), green dashed line], the occupancy is off since *in vivo* potential wells cannot be captured by this model [Fig. 4(c), light blue dash-dot line]. Nucleosomes are not strongly localized if the *in vivo* wells and barriers are absent [Fig. 4(c), blue solid line], consistent with the relatively smooth *in vitro* occupancy profiles [12,13]. Note that the two occupancy profiles will coincide if the mean of the predicted one-body energies is set to the correct value, eliminating the spurious offset caused by Φ [Fig. 4(a), cf. blue solid and light blue dash-dot lines].

In order to reconstruct the occupancy correctly and avoid mixing one-body and two-body contributions, we need to turn to the full theory developed in section 2.1. Inserting predicted Φ [Fig. 4(b), green dashed line] into Eq. (3) and using the exact $n(i)$ profile, we obtain a system of nonlinear equations which can be solved numerically, yielding energies that are very close to the exact result [Fig. 4(a), maroon dotted line]. These energies and the predicted Φ can be used to reconstruct the occupancy profile which is nearly exact [Fig. 4(c), cf. green dash-dot-dot and maroon dotted lines]. Thus we have succeeded in separating one- and two-body effects, and in splitting off the sequence-dependent part in the former. However, the full procedure is computationally intensive and becomes inefficient if the DNA is longer than 10^4 bp (longer segments may be split into manageable pieces and handled separately).

3.2 Nucleosome localization by potential barriers and wells

Nucleosomes in the vicinity of potential barriers and wells can be localized by steric exclusion alone [17]. This mechanism is thought to contribute to prominent nucleosome occupancy peaks in genic regions observed *in vivo* but not *in vitro* [12,13,14,15,16,21]. In order to understand the nature and the extent of *in vivo* nucleosome localization, we need to study nucleosome occupancy patterns created by placing a single potential barrier or potential well onto an otherwise flat one-body energy landscape.

In Fig. 5 we show nucleosome occupancy induced by a symmetric Gaussian barrier, with and without two-body interactions. As the chemical potential is changed to increase the average occupancy, the
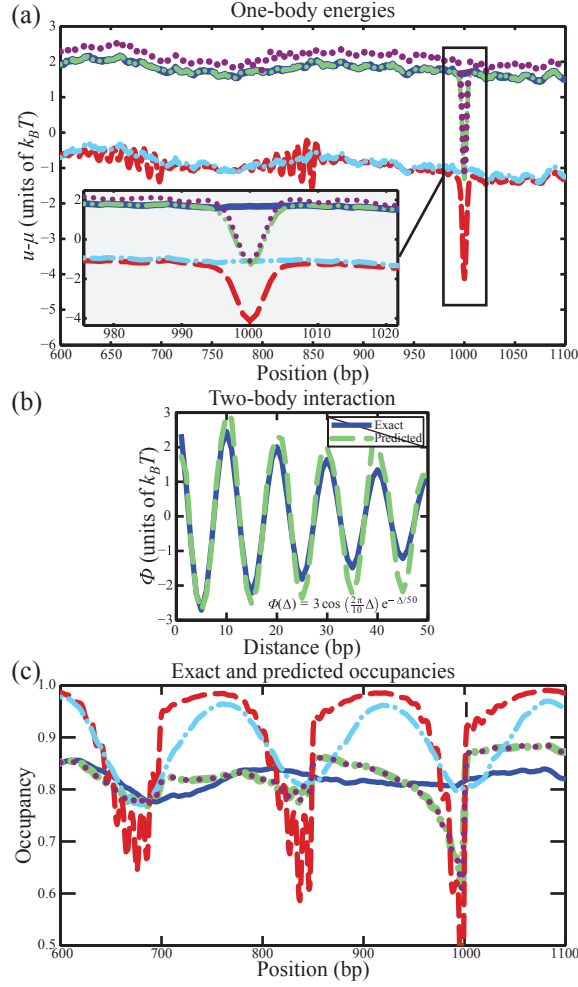
Fig. 4 (Color online) (a) One-body energies. Sequence-dependent energy given by the 21-parameter model [Eq. (26)] (blue solid line), total energy given by the sum of the sequence-specific energies and 5 potential wells with 3 k_BT depth at $1, 3, 5, 7$, and 9×10^3 bp designed to mimic the *in vivo* effects (green dash-dot-dot line). Energy predicted with a model that neglects two-body interactions [Eq. (10)] (red dashed line), energy predicted by fitting the 21-parameter model to the energies from Eq. (10) (light blue dash-dot line), a numerical solution of the full model which takes Φ into account (maroon dotted line). Inset: zoom-in on the region with the potential well. (b) Exact two-body interaction Φ (blue solid line) and predicted interaction [Eq. (17)] (green dashed line). (c) Nucleosome occupancies. Occupancy generated by the exact sequence-specific one-body energy and the exact interaction (blue solid line), occupancy corresponding to the combined exact one-body energy (sequence-specific component and potential wells) and the exact interaction (green dash-dot-dot line). Predicted occupancy generated by the one-body energy from Eq. (10) and predicted Φ (red dashed line), occupancy generated using predicted sequence-dependent one-body energy [Eq. (26)] and predicted Φ (light blue dash-dot line), occupancy predicted using numerically computed one-body energies from the full model and predicted Φ (maroon dotted line).

oscillations become more prominent. Without two-body interactions, the peak closest to the barrier is always the highest and the occupancy pattern is a decaying oscillation [Fig. 5(a)]. Strikingly, including Φ results in a markedly different occupancy profile: oscillations are more persistent and the peak closest to the barrier is not always the highest [Fig. 5(b)].

The degree of nucleosome localization is also controlled by the width of the Gaussian barrier: wider barriers induce less prominent oscillations, but produce stronger occupancy depletion over the barrier itself [Figs. 5(c) and 5(d)]. The degree of depletion is also controlled by the barrier height [Figs. 5(e) and 5(f)]. Interestingly, increasing the strength of two-body interactions results in a higher average occupancy and produces shorter peak-to-peak distances [Fig. 5(g)]. In fact, the peak-to-peak distances
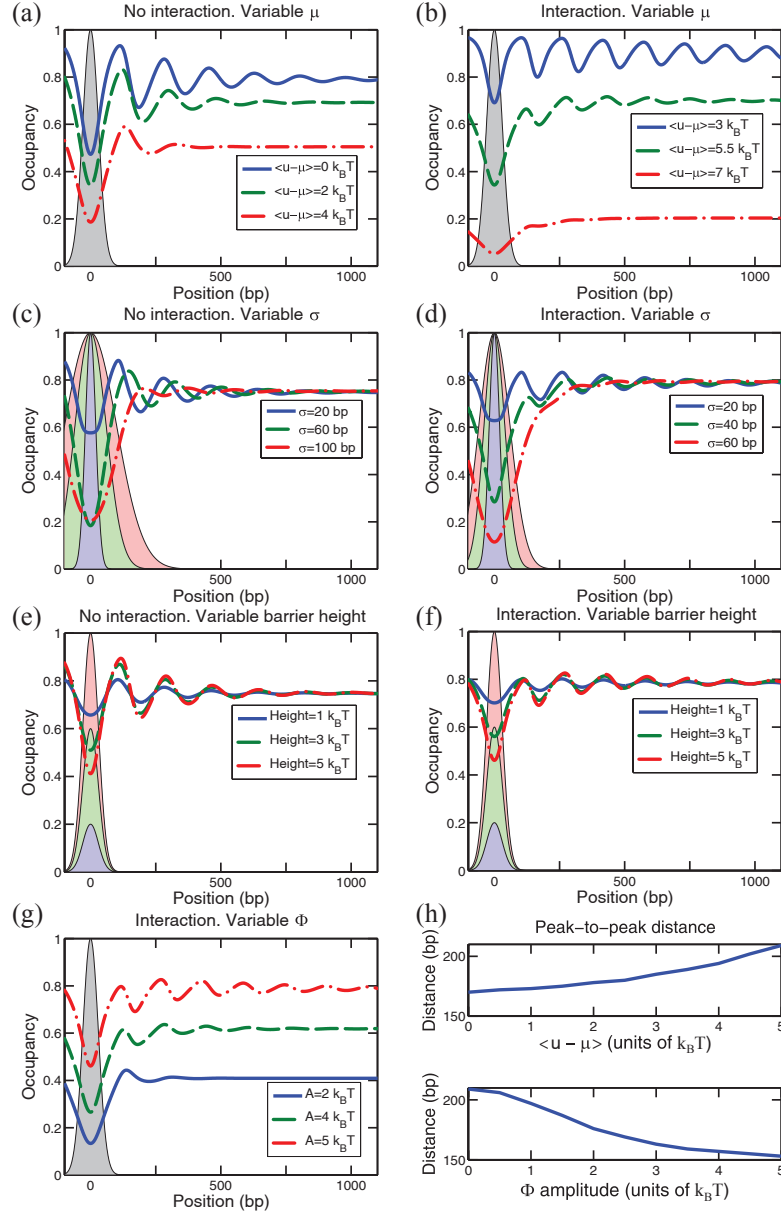
Fig. 5 (Color online) Symmetric Gaussian barrier. Occupancy profiles for the following scenarios: variable chemical potential μ (a) and (b), variable barrier width (c) and (d), and variable barrier height (e) and (f). Unless otherwise specified in the legend, the barrier heights are $5\ k_BT$, $\sigma = 30$ bp, and $\langle u-\mu \rangle = 5\ k_BT$ [in panel (c) $\langle u-\mu \rangle = 1\ k_BT$]. Panels (b), (d) and (f) have a two-body interaction $\Phi(\Delta) = A\cos(2\pi\Delta/10)\exp(-\Delta/50)$, with $A = 5\ k_BT$. (g) Occupancy profiles for variable interaction strength A. (h) Variation of the typical distance between neighboring nucleosomes as μ or A is varied. Upper panel: $\Phi = 0$, lower panel: $\langle u \rangle - \mu = 5\ k_BT$.

(which can be interpreted as the sum of the 147 bp nucleosomal site and a linker) can be varied in a wide range by changing either the chemical potential μ or the strength of Φ [Fig. 5(h)].

Similar conclusions can be reached if the Gaussian barrier is replaced by a symmetric Gaussian potential well (Fig. 6): oscillations decay less rapidly with two-body interactions, and the extent of oscillations is controlled by the chemical potential and by the depth and the width of the well. However, in this case the nucleosome closest to the well is always the most localized. Nucleosome occupancy in the vicinity of 5' NDRs is prominently asymmetric [13,16]. This asymmetry can be modeled by a combination of a symmetric barrier with an adjacent potential well [27], or by a single asymmetric
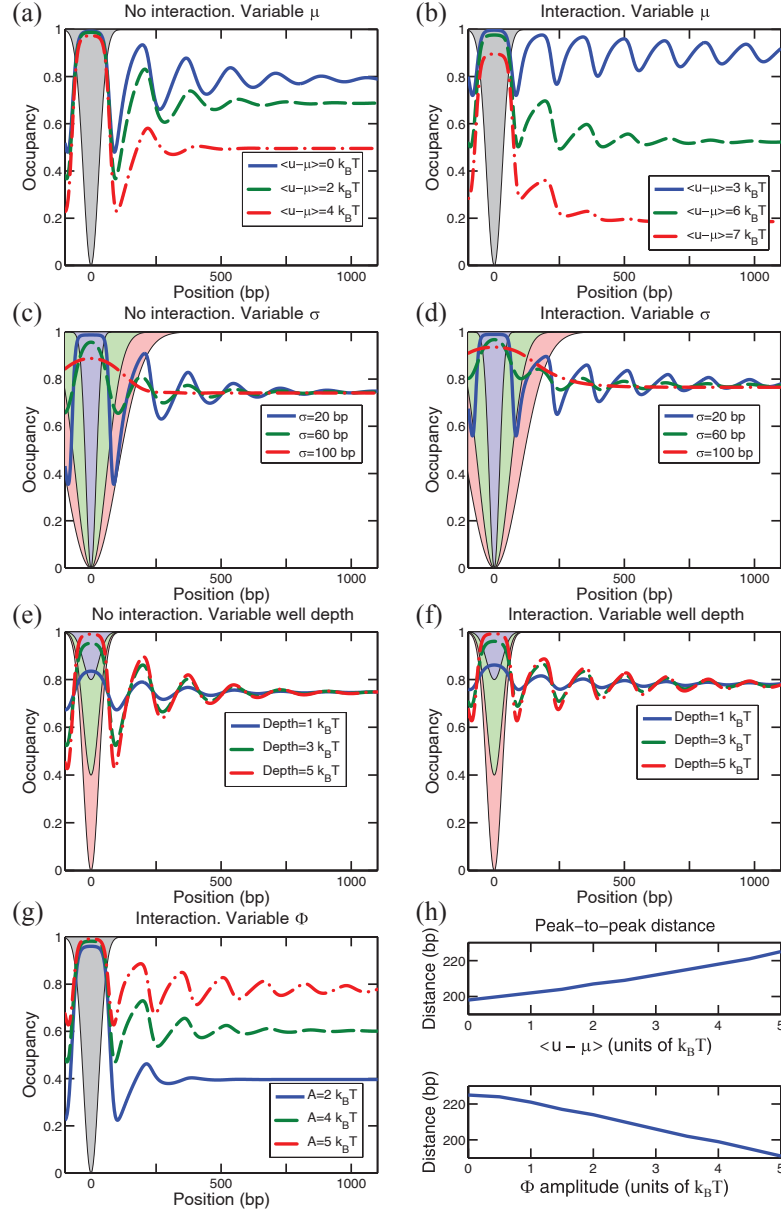
Fig. 6 (Color online) Symmetric Gaussian well. Occupancy profiles for the scenarios described in Fig. 5. All the parameters not explicitly given in the legends are from Fig. 5. In particular, well depths have the same magnitude as the heights of the corresponding barriers.

barrier. In Fig. 7 we show how nucleosome localization and the degree of asymmetry in the occupancy profile vary with the chemical potential, the strength of the two-body interaction, the height of the barrier, and the degree of its asymmetry.

In summary, two-body interactions significantly modify nucleosome occupancy profiles, affecting heights and spacings of observed nucleosome localization peaks. However, as the interaction itself only couples neighboring nucleosomes but does not determine their absolute positions, a potential barrier or well is required to achieve localization in the first place. Increasing the width of this feature diminishes its localization capacity. The interaction favors configurations with linker lengths corresponding to the minima of Φ, leading to linker length discretization [25,26,27]. By changing the interaction strength or the chemical potential one can create occupancy patterns with different average linker lengths.
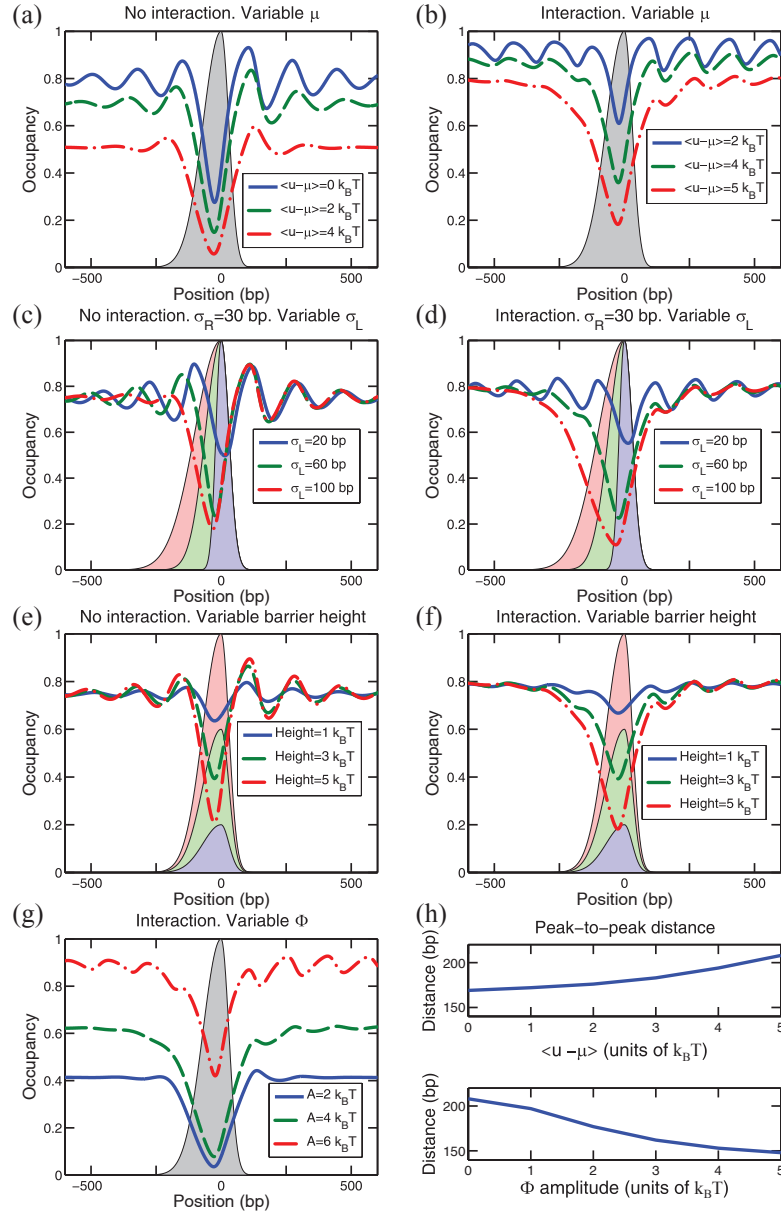
Fig. 7 (Color online) Asymmetric Gaussian barrier. Occupancy profiles for the scenarios described in Fig. 5. Unless otherwise specified in the legend, the barrier heights are 5 $k_B T$, $\sigma_L = 70$ bp, $\sigma_R = 30$ bp, and $\langle u - \mu \rangle = 5$ $k_B T$ [in panel (c) $\langle u - \mu \rangle = 1$ $k_B T$].

## 3.3 Modeling nucleosome occupancy over transcribed regions

The characteristic patterns of nucleosome occupancy in the region between 5' and 3' NDRs are shown in Fig. 8. As discussed in the Introduction, there is a pronounced lack of nucleosome localization *in vitro* [12,13] [Figs. 8(a) and 8(b)]. The 21-parameter $N = 2$ position-independent model captures this liquid-like behavior correctly but is unable to account for the *in vivo* peaks. Since DNA sequence alone clearly cannot produce the observed degree of *in vivo* localization, we sought to construct a minimal model in which potential barriers of non-sequence origin flank each gene and the one-body energy landscape is flat otherwise [37,38] (Fig. 9).

In the Kaplan et al. dataset [12], the first nucleosome is in fact the most localized and the average profile is consistent with the absence of two-body interactions [Fig. 8(a)]. In contrast, Zawadzki et
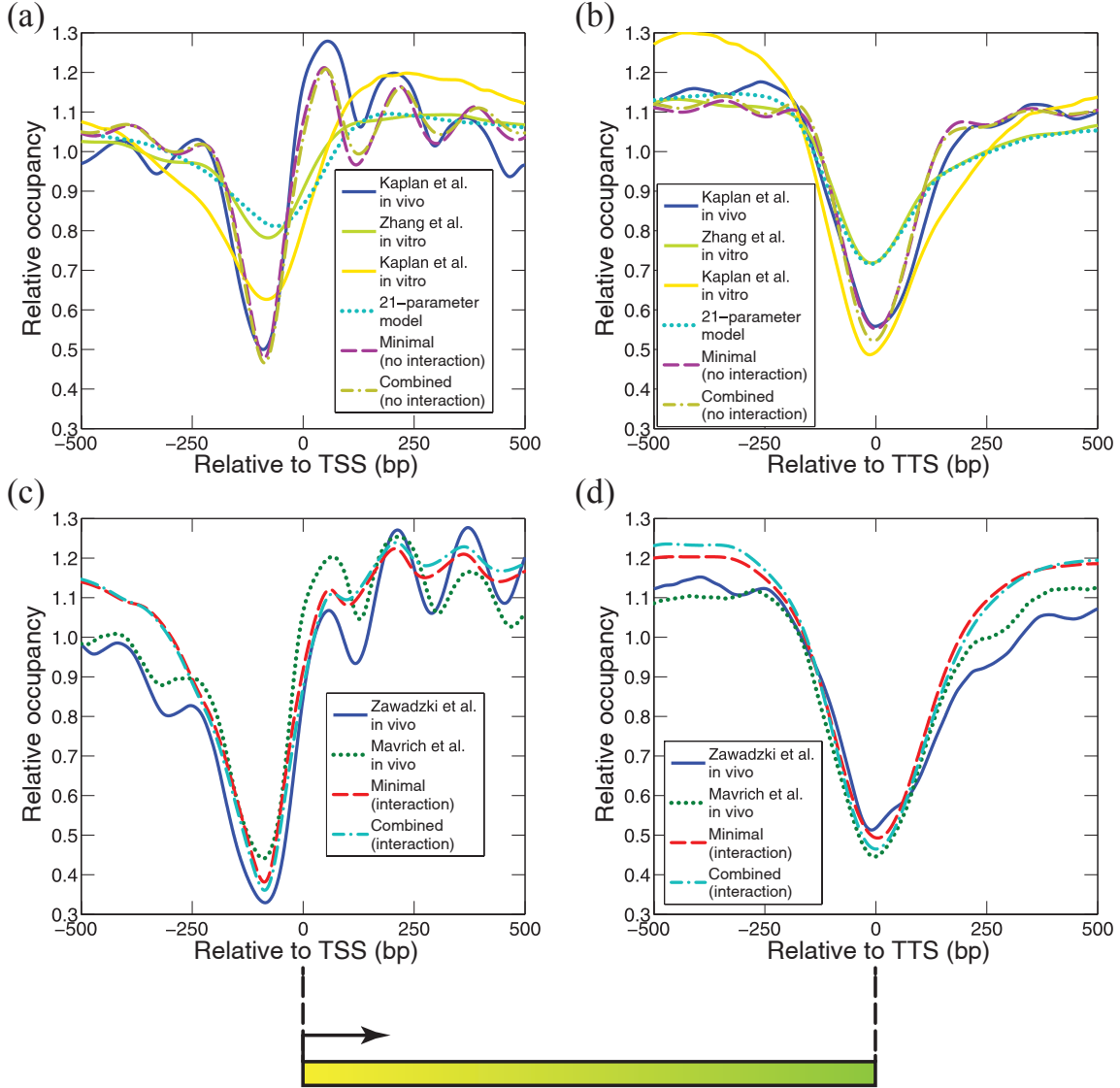
Fig. 8 (Color online) Average nucleosome occupancy in the vicinity of transcription start and termination sites (TSS and TTS, respectively). Each occupancy profile is normalized by its average in the $[-500, 500]$ bp window. (a), (b): Nucleosome occupancy observed *in vivo* (YPD medium) and *in vitro* by Kaplan et al. [12] and *in vitro* by Zhang et al. [13], and predicted using a 21-parameter $N = 2$ position-independent model, a minimal model in which nucleosomes are localized purely by means of sequence-independent potential barriers (Fig. 9), and a combined model in which sequence-specific energies from the 21-parameter $N = 2$ model are added to the barriers from Fig. 9. The two-body potential is turned off. Note that in Ref. [13] DNA was mixed with histones in a 1:1 mass ratio which is close to the *in vivo* value, while in Ref. [12] the ratio was 0.4:1, resulting in deeper NDRs. (c), (d): Nucleosome occupancy observed *in vivo* by Zawadzki et al. [21] and Mavrich et al. [16], and predicted using the 21-parameter $N = 2$ position-independent model, the minimal model, and the combined model. The two-body potential is given by $\Phi(\Delta) = A \cos(2\pi\Delta/10) \exp(-\Delta/50)$, with $A = 5 \, k_B T$.

al. [21] and Mavrich et al. [16] profiles appear to be shaped by the higher-order chromatin structure [Fig. 8(c)]. This experimental discrepancy may have resulted from under-digesting chromatin with micrococcal nuclease (MNase), an enzyme typically used to liberate mononucleosome cores [39]. In addition, the number of active genes that presumably reside in more open, active chromatin characterized by weaker two-body interactions could vary between experiments. However, in all three cases the *in vivo* barriers are necessary to reproduce observed localization patterns. The 5' NDR is strongly
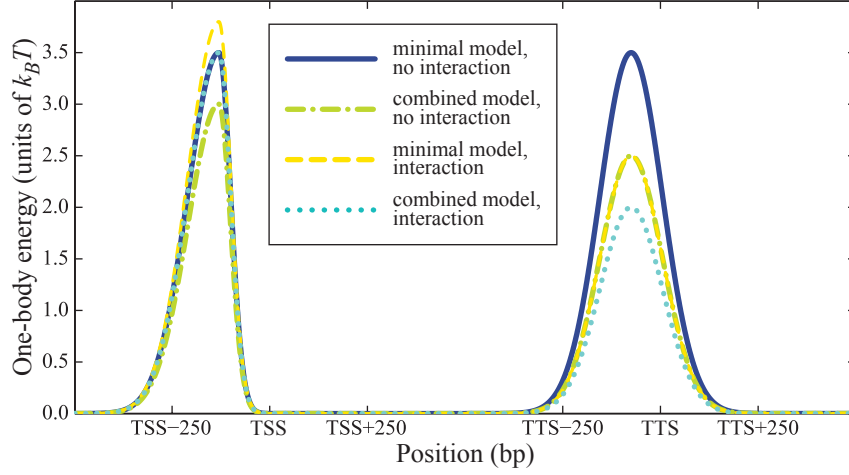
Fig. 9 (Color online) The one-body energy profiles used in Figs. 8 and 10. The 5' asymmetric barrier has $\sigma_{\text{left}} = 80$ bp and $\sigma_{\text{right}} = 30$ bp. The 3' symmetric barrier has $\sigma = 80$ bp. Solid blue line: barriers used in the *in vivo* minimal model without two-body interactions [Figs. 8(a) and 8(b), Fig. 10]. Dash-dot green line: barriers used in the *in vivo* combined model without two-body interactions [Figs. 8(a) and 8(b), Fig. 10]. Dashed yellow line: barriers used in the *in vivo* minimal model with two-body interactions [Figs. 8(c) and 8(d)]. Dotted light blue line: barriers used in the *in vivo* combined model with two-body interactions [Figs. 8(c) and 8(d)]. The landscapes shown in the Figure are shifted vertically so that $\langle u - \mu \rangle = 0.56\ k_B T$ in the minimal model without two-body interactions, $0.62\ k_B T$ in the combined model without two-body interactions, $4.49\ k_B T$ in the minimal model with two-body interactions, and $4.62\ k_B T$ in the combined model with two-body interactions.

asymmetric [Figs. 8(a) and 8(c)] and thus needs to be modeled either with a combination of a symmetric barrier and a potential well for the +1 nucleosome [27], or with a single asymmetric barrier (Fig. 7). The height of each barrier in Fig. 9 is adjusted to reproduce the extent of observed nucleosome depletion.

The average occupancy profiles are not significantly altered if sequence-specific energies from the 21-parameter $N = 2$ position-independent model are added to the barriers from Fig. 9 (Fig. 8, cf. combined and minimal models). The $N = 2$ model yields a standard deviation of $0.61\ k_B T$ for the energies genome-wide, consistent with the assumption that sequence-dependent energies should be less than $1\ k_B T$, and thus the one-body landscape is still dominated by the barriers. Note that barrier heights are reduced in the combined model because sequence-dependent nucleosome depletion over NDRs is now included explicitly.

The difference between the minimal and the combined models is more pronounced if individual occupancy profiles are displayed as a heat map (Fig. 10). Minimal model barriers adjacent to each other on the genomic sequence (*e.g.* the 5' barriers of two divergent genes sharing a single promoter) sometimes create anomalous NDRs with the extent of nucleosome depletion not observed in the data [Figs. 10(e) and 10(f)]. Interestingly, these effects are reduced when sequence specificity is combined with the minimal model [Figs. 10(g) and 10(h)]. Comparing barrier heights in the minimal and combined models (Fig. 9), we conclude that intrinsic histone-DNA interactions are responsible for $< 30\%$ of the barriers.

4 Discussion and Conclusion

We have developed a theory of nucleosomes positioned by both intrinsic histone-DNA interactions and a two-body, nearest-neighbor potential induced by higher-order chromatin structure. Our theory is general and can be applied to any system of finite-size particles in an external field of arbitrary strength, interacting via a short-range two-body potential. From one- and two-body nucleosome energies we can reconstruct one- and two-particle distribution functions exactly. Conversely, one- and two-particle distributions available from experiment can be used to predict both intrinsic histone-DNA interactions and the two-body potential with high accuracy, even in strongly inhomogeneous systems dominated
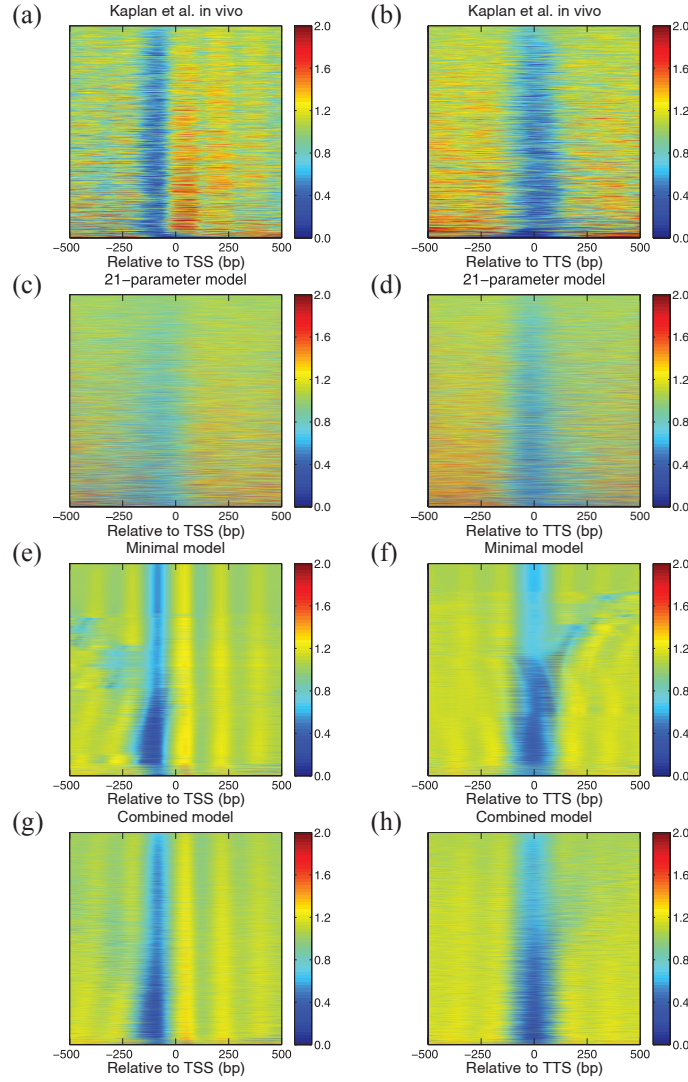
Fig. 10 (Color online) Heat maps of nucleosome occupancy around TSS and TTS for 5747 *S. cerevisiae* genes. *In vivo* nucleosomes (YPD medium) [12] (a) and (b), $N = 2$ position-independent model (c) and (d), minimal model (e) and (f), combined model (g) and (h). The minimal model is constructed by placing potential barriers from Fig. 9 at the end of each gene onto an otherwise flat one-body energy landscape without two-body interactions. The combined model is constructed by adding sequence-specific energies from the 21-parameter $N = 2$ position-independent model (which have standard deviation of 0.61 $k_B T$ genome-wide) to the minimal model. The occupancy for each gene is normalized by the average occupancy in the $[-500, 500]$ bp window. The experimental data [(a) and (b)] are smoothed with a $2D$ Gaussian kernel ($\sigma_X = 1$ bp and $\sigma_Y = 2$ genes). The genes are sorted in each panel in the order of increasing variance of the occupancy. The genome-wide average occupancies are 0.1508 [(a) and (b)], 0.2024 [(c) and (d)], 0.7516 [(e) and (f)], and 0.7232 [(g) and (h)].

by one-body forces. Finally, we infer sequence determinants of nucleosome positions through a linear fit of predicted one-body energies to a model in which each mono- and dinucleotide is assigned the same energy regardless of its position within the nucleosomal site (the $N = 2$ position-independent model). In contrast to our previous approach which employed a hierarchy of energy constraints [22], here we chose to fit all available parameters. The fitted parameters are subsequently projected onto the relevant subspace, creating a non-degenerate description of nucleosome energetics.

Unfortunately, high-throughput maps of nucleosome positions typically yield just one-particle distributions – two-body information is lost because the positional data is averaged over many cells. Nonetheless, the two-body potential can be inferred from one-body density profile $n(i)$ if Φ is suffi-

ciently strong to create $10 - 11$ bp periodic occupancy oscillations in the vicinity of potential wells and barriers which tend to localize nucleosomes (Fig. 4). These oscillations can then be used to predict P_{linker}, the distribution of linker lengths, which gives Φ via Eq. (17). This empirical approach fails in the absence of one-body energies because arrays of nucleosomes (even subject to two-body forces) can freely move along the DNA without any energy cost, resulting in a constant $n(i)$ from which P_{linker} cannot be deduced.

We have previously shown that our approach is capable of separating the $10 - 11$ bp periodic two-body potential from the one-body contribution which is responsible for rotational nucleosome positioning and thus has the same periodicity of DNA helical twist [27]. Furthermore, we have inferred two-body interactions from genome-wide maps of nucleosomes assembled *in vitro* on genomic DNA, and have demonstrated their essential role in explaining the observed autocorrelation of genome-wide nucleosome positions, and in shaping the occupancy profile over transcribed regions.

However, the full analysis, which involves finding $u(i)$ from a system of nonlinear equations obtained by inserting Φ from Eq. (17) into Eq. (3), is computationally expensive. A simpler approach which neglects two-body interactions [Eq. (10)] is much faster but leads to "leakage" of two-body effects into predicted one-body potentials if Φ is in fact present in the system (Fig. 4). Nonetheless, if a sequence-specific model is fit to this one-body profile, the correct sequence-dependent energy is obtained because oscillations caused by Φ are not related to any sequence features. Thus if one only needs to determine sequence determinants of nucleosome positioning, the two-body potential can be neglected altogether and a simpler, computationally efficient approach can be used. Nevertheless, only the full theory presented here is capable of disentangling one- and two-body effects.

In vivo, nucleosomes adjacent to 5' and 3' NDRs are positioned by steric exclusion which creates prominent localized peaks of nucleosome occupancy in the vicinity of each depleted region [12, 14, 15, 16]. This localization is absent *in vitro* [12, 13], indicating that intrinsic histone-DNA interactions can only partially account for *in vivo* occupancy patterns, creating barriers with heights less than $1 \, k_B T$. Therefore, sequence-specific models need to be combined *in vivo* with additional potentials established by RNA polymerase, ATP-dependent chromatin remodeling enzymes, and DNA-binding proteins.

To this end, we have studied nucleosome positioning by potential wells and barriers, showing in particular that the distance between neighboring peaks in the nucleosome occupancy profile depends on both histone octamer concentration and the strength of two-body interactions (Figs. 5, 6 and 7). We then created a genome-wide model of nucleosome occupancy by flanking each transcribed region with potential barriers whose height, width and the degree of asymmetry were adjusted to reproduce the observed occupancy patterns (Fig. 9). This minimal, sequence-independent model successfully captures the main features of *in vivo* occupancy, although, surprisingly, the role of Φ seems to vary between experiments (Fig. 8). When the minimal model is combined with the sequence-specific energies (with magnitudes $\leq 1 \, k_B T$ as discussed above), occupancy patterns of a subset of genes get closer to experiment (Fig. 10), but the effect of DNA sequence on the average occupancy profile is rather small (Fig. 8).

In summary, we have presented a rigorous theory of nucleosomes subject to nearest-neighbor two-body interactions, demonstrating the essential role of chromatin structure in genome-wide models of nucleosome occupancy. Future experiments focused on multi-nucleosome distributions will provide data for our exact theory [Eqs. (8) and (9)], obviating the need for the approximate Eq. (17). Further experimental and computational work should also be aimed at pinpointing the origin of *in vivo* potential barriers in various organisms and cell types.

References

1. G. Felsenfeld, M. Groudine, Nature (London) **421**, 448 (2003)
2. T.J. Richmond, C.A. Davey, Nature (London) **423**, 145 (2003)
3. K. Struhl, Cell **98**(1), 1 (1999)
4. J. Mellor, Trends Genet. **22**(6), 320 (2006)
5. B. Li, M. Carey, J.L. Workman, Cell **128**(4), 707 (2007)
6. L. Bai, A.V. Morozov, Trends Genet. **26**(11), 476 (2010)
7. W.K. Olson, et al., Proc. Natl. Acad. Sci. U.S.A. **95**(19), 11163 (1998)

8. A.V. Morozov, et al., Nucleic Acids Res. **37**(14), 4707 (2009)
9. P.T. Lowary, J. Widom, J. Mol. Biol. **276**(1), 19 (1998)
10. A. Thastrom, et al., J. Mol. Biol. **288**(2), 213 (1999)
11. E.A. Sekinger, Z. Moqtaderi, K. Struhl, Mol. Cell **18**(6), 735 (2005)
12. N. Kaplan, et al., Nature (London) **458**, 362 (2009)
13. Y. Zhang, et al., Nature Struct. Mol. Biol. **16**(8), 847 (2009)
14. G.C. Yuan, et al., Science **309**(5734), 626 (2005)
15. T.N. Mavrich, et al., Nature (London) **453**, 358 (2008)
16. T.N. Mavrich, et al., Genome Res. **18**, 1073 (2008)
17. R.D. Kornberg, L. Stryer, Nucleic Acids Res. **16**(14), 6677 (1988)
18. J.L. Workman, Genes Devel. **20**, 2009 (2006)
19. L.O. Barrera, B. Ren, Curr. Opin. Cell Biol. **18**(3), 291 (2006)
20. G.R. Schnitzler, Cell Biochem. Biophys. **51**(2-3), 67 (2008)
21. K.A. Zawadzki, A.V. Morozov, J.R. Broach, Mol. Biol. Cell **20**(15), 3503 (2009)
22. G. Locke, et al., Proc. Natl. Acad. Sci. U.S.A. **107**(49), 20998 (2010)
23. D. Tolkunov, A. Morozov, Adv. Protein Chem. Struct. Biol. **79**, 1 (2010)
24. L.E. Ulanovsky, E.N. Trifonov, *Biomolecular Stereodynamics III* (Adenine Press, Schenectady, NY, 1986), pp. 35–44
25. J. Widom, Proc. Natl. Acad. Sci. USA **89**(3), 1095 (1992)
26. J.P. Wang, et al., PLoS Comput. Biol. **4**(9), e1000175 (2008)
27. R.V. Chereji, D. Tolkunov, G. Locke, A.V. Morozov, Phys. Rev. E **83**(5), 050903 (2011)
28. H. Takahashi, in *Proc. Phys.-Math. Soc. Japan*, vol. 24 (1942), vol. 24, p. 60
29. H.L. Frisch, J.L. Lebowitz (eds.), *The Equilibrium Theory of Classical Fluids* (W. A. Benjamin Inc., New York, NY, 1964)
30. J.K. Percus, J. Stat. Phys. **15**(6), 505 (1976)
31. J.K. Percus, J. Phys.: Condens. Matter **1**(17), 2911 (1989)
32. F. Gürsey, in *Math. Proc. Cambridge Philos. Soc.*, vol. 46 (1950), vol. 46, pp. 182–194
33. Z. Salsburg, R. Zwanzig, J. Kirkwood, J. Chem. Phys. **21**(6) (1953)
34. I.Z. Fisher, T.M. Switz, *Statistical theory of liquids* (University of Chicago Press, Chicago, IL, 1964)
35. L. Tonks, Phys. Rev. **50**(10), 955 (1936)
36. E. Segal, et al., Nature (London) **442**, 772 (2006)
37. W. Möbius, U. Gerland, PLoS Comput. Biol. **6**(8), e1000891 (2010)
38. G. Chevereau, et al., Phys. Rev. Lett. **103**(18), 188103 (2009)
39. A. Weiner, et al., Genome Res. **20**, 90 (2010)